AD-A054 886 DELAWARE UNIV NEWARK DEPT OF STATISTICS AND COMPUTER--ETC F/G 12/1
ON THE TRADE-OFF BETWEEN QUEUE CONGESTION AND SERVER'S REWARD I--ETC(U)
APR 78 G LATOUCHE AFOSR-77-3236

UNCLASSIFIED TR-78/7 AFOSR-TR-78-0985 NL

| OF |
AD
A054886

END
DATE
FILMED
7-78
DDC

Department of
STATISTICS AND COMPUTER SCIENCE

UNIVERSITY OF DELAWARE
Newark, Delaware 19711

(2)

ON THE TRADE-OFF BETWEEN QUEUE CONGESTION AND SERVER'S REWARD IN A M/M/1 QUEUE.

by

(10) Guy Latouche
Université Libre de Bruxelles
and
University of Delaware

(16) 2304 (17) A5

(18) AFOSR (19) TR-78-0925

Department of Statistics
and
Computer Science

(9) Technical Report, No. (14) TR-78/7

(11) April 1978 (12) 34p.

(15) This research was sponsored by the Air Force Office of Scientific Research Air Force Systems Command USAF, under ~~Grant No.~~ AFOSR-77-3236. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

1473

391 807 LB

ABSTRACT

We consider an M/M/1 queue which is controlled by dynamically setting the customer's entrance fee to either b_1 or $b_2 > b_1$ and thereby setting the customer arrival rate to either λ_1 or $\lambda_2 < \lambda_1$, respectively. With E[F] the expected fee collected per unit of time and P_N the steady state probability that the system contains more than N customers, we consider two criteria: (i) for some number $\bar{r}$, minimize P_N subject to $E[F] \geq \bar{r}$, and (ii) for some number ε, maximize E[F] subject to $P_N \leq \varepsilon$. Under each criterion we consider two cases: (a) the admissible policies are single critical number switching policies and (b) a cost $\gamma \geq 0$ is incurred whenever the server switches between b_1 and b_2, the admissible policies allow for hysteresis to appear in the arrival process. Optimal policies are computed for each criterion and each case.

One can observe, in the literature on queueing theory, a growing interest in the control of the arrival process. For most of the published models, the aim is to maximize the difference between fees collected and costs: it is explicitly assumed that the fees and costs can be measured and are expressed in the same units. They are then merged into a single objective function (Naor [7], Yechiali [12,13], Edelson & Hildebrand [1], Teghem [10] or Low [5,6], among others.)

In fact, it is recognized that such a merging cannot always be done. Just to give one example, we observe that the fee charged to customers may be an arbitrary notion (as noted by Nielsen [8]: "the pricing of computer services is not dependent upon charging users real money"); in such a case, it might be difficult to express costs in the same unit as the fee. We assume in this paper that the reason for controlling the arrival process is to achieve some balance between acceptance and rejection of customers; while letting customers in increases the fee collected, an excess of customers should be avoided.

In this paper we consider the two objectives of maximizing the fee collected and minimizing queue congestion as distinct. If it is not possible to optimize these simultaneously, then the problem will be to maintain one of them at a reasonable level and, under that constraint, to optimize the other.

In another paper [4], we considered similar models where the two objectives are merged into one single function. The analysis and results in [4] are both quantitatively and qualitatively different from those we present here.

Specifically, we assume that the server may dynamically set the entrance fee to either b_1 or $b_2 > b_1$; the arrivals form a Poisson process with parameter λ_1 or $\lambda_2 < \lambda_1$ respectively.

The service times are independent, identically distributed, exponential (μ) random variables, independent of the arrival process.

The waiting room is infinite, but the total number of customers in the system should preferably not exceed some value N, called the critical level.

With E[F], the expected fee collected per unit of time, in steady state and P_N the steady state probability that the system contains more than N customers, we consider two criteria:

(i) for some number $\bar{r}$ minimize P_N, subject to $E[F] \geq \bar{r}$, and

(ii) for some number ε maximize E[F], subject to $P_N \leq \varepsilon$.

Under each criterion, we consider two cases:

(a) the admissible policies are single critical number switching policies, they are characterized by a number M: if there are fewer than M customers in the system, the entrance fee is b_1, otherwise it is b_2;

(b) a cost $\gamma \geq 0$ is incurred whenever the server switches between b_1 and b_2 and the admissible policies allow for hysteresis to appear in the arrival process. These policies are characterized by two numbers, m and M ($m \leq M-1$): if the fee is b_1 and the number of customers increases from (M-1) to M, the fee becomes b_2; if the fee is b_2 and the number of customers decreases from (m+1) to m, the fee becomes b_1.

Hysteresis in the service process has been considered by several authors (for instance, Yadin & Naor [11] or Teghem [10]). To the best of our knowledge, only Scott [9] has considered hysteresis in the arrival process; we comment more on this paper in Section 4.

Optimal policies are computed for each criterion and each case.

The policies we determine are not necessarily optimal with respect to larger sets of admissible control policies. For instance, we mention in Sections 6 and 7 that even with $\gamma=0$, some policy with hysteresis may be strictly superior to every single critical number policy. Moreover, one can determine randomized stationary policies which are strictly superior to any policy in the classes we have considered. Nevertheless, these policies are easily implemented and can be efficient, as we show by examples in Section 8.

The optimal single critical number policy is determined explicitly for the first criterion; for the second criterion, the optimal policy is determined explicitly for large ε, through an equation with a unique solution for small ε (this is made precise in Section 3). To find the optimal policy with hysteresis, we use properties of $E[F]$ and P_N to determine a finite number of pairs (m,M) among which the best is to be found numerically. The computations involved are straightforward. We shall not present detailed proofs; some of them are quite lengthy and are presented in [3].

To conclude the introduction, we comment on our measure of queue congestion, the probability P_N of having more than N customers in the system. In some cases, a more appropriate measure might be the expected queue length or the mean cost of waiting if customers were able to measure such a cost.

For the first criterion, the optimal single critical number policy would be the same; it appears in Section 2 that the optimal value for M does not depend on N. In the other cases, we suspect that the analysis would follow the same lines although the details are of course different.

1. Control With a Single Critical Number

In this case, the model is simple to analyze and most of the results in Sections 1 to 3 are intuitively obvious. As stated in the introduction, we assume $b_1<b_2$ and $\lambda_1>\lambda_2$.

Moreover, if $\rho_i=\frac{\lambda_i}{\mu}$, $i=1,2$, we assume that $\rho_2<1$, while ρ_1 may be less than or greater than 1.

It is obvious, therefore, that steady state conditions are satisfied if $M<\infty$ or $M=\infty$ and $\rho_1<1$.

Let

$$\pi_i(M) = P[\text{i customers in the system in steady state}|M]$$

and

$$P_i(M) = \sum_{j=i+1}^{\infty} \pi_i(M),$$

we shall refer to the P_i's as the tail probabilities.

One can easily solve the system of equations for the probabilities $\pi_i(M)$ $(i=0,1,\ldots)$ and, from there, one gets for $\rho_1\neq 1$

$$P_i(M) = \frac{\rho_1^{i+1}(1-\rho_2)-\rho_1^M(\rho_1-\rho_2)}{1-\rho_2-\rho_1^M(\rho_1-\rho_2)} \qquad i\leq M, \tag{1a}$$

$$\frac{(1-\rho_1)\rho_1^M\rho_2^{i-M+1}}{1-\rho_2-\rho_1^M(\rho_1-\rho_2)} \qquad i\geq M; \tag{1b}$$

for $\rho_1=1$, the corresponding formulas are

$$P_i(M) = \frac{1+(M-i-1)(1-\rho_2)}{1+M(1-\rho_2)} \qquad i\leq M,$$

$$= \frac{\rho_2^{i-M+1}}{1+M(1-\rho_2)} \qquad i\geq M.$$

Lemma 1: For all i, $P_i(M)$ is an increasing function of M.

This property is intuitively obvious: if M increases, the number of states for which the arrival rate is λ_1 ($>\lambda_2$) increases and the probability of having more than i customers increases. Now, let

$r(M)$ = E[fee collected per unit of time in steady state|M].

Of course,

$$r(M) = \lambda_2 b_2 + (\lambda_1 b_1 - \lambda_2 b_2)(1 - P_{M-1}(M)). \tag{2}$$

where

$$P_{M-1}(M) = \frac{\rho_1^M(1-\rho_1)}{1-\rho_2-\rho_1^M(\rho_1-\rho_2)} \quad .$$

Lemma 2: $r(M)$ is an increasing function of M if $\lambda_1 b_1 > \lambda_2 b_2$; it is constant if $\lambda_1 b_1 = \lambda_2 b_2$ and decreasing if $\lambda_1 b_1 < \lambda_2 b_2$.

This results from the fact that $P_{M-1}(M)$ is decreasing in M for all finite value of ρ_1. The following result is an immediate consequence of Lemmas 1 and 2:

Theorem 1: If $\lambda_1 b_1 \leq \lambda_2 b_2$, it is optimal to set M equal to 0.

In fact, only if $\lambda_1 b_1 \leq \lambda_2 b_2$ can one reach simultaneously the two objectives: maximize $r(M)$ and minimize $P_N(M)$ and it is done by choosing the fee b_2 always. We shall assume in the next two sections that $\lambda_1 b_1 > \lambda_2 b_2$.

2. Constraint on the Expected Fee

Suppose it is decided that the expected fee r(M) must be at least equal to some predetermined value $\bar{r}$. The problem is then to find M_1 such that

$$r(M_1) \geq \bar{r} \tag{3}$$

and

$$P_N(M_1) \leq P_N(M) \text{ for all } M \text{ such that } r(M) \geq \bar{r}.$$

Let $r(\infty) = \lim_{M\to\infty} r(M)$, one gets easily from (1) and (2) that

$$\lim_{M\to\infty} P_{M-1}(M) = 0 \qquad \text{if } \rho_1 \leq 1,$$

$$= \frac{\rho_1 - 1}{\rho_1 - \rho_2} \qquad \rho_1 > 1,$$

hence

$$r(\infty) = \lambda_1 b_1 \qquad \rho_1 \leq 1,$$

$$= \lambda_2 b_2 \frac{\rho_1 - 1}{\rho_1 - \rho_2} + \lambda_1 b_1 \frac{1 - \rho_2}{\rho_1 - \rho_2} < \lambda_1 b_1 \qquad \rho_1 > 1.$$

Clearly, if M is equal to ∞, the expected fee over any interval of time is $\lambda_1 b_1$ times the length of that interval, even if $\rho_1 > 1$. Therefore, if $r(M=\infty)$ denotes the value of r(M) when $M=\infty$, we shall consider that $r(M=\infty) = \lambda_1 b_1$ for all λ_1.

Theorem 2:

If	$\lambda_1 b_1 < \bar{r}$, then	there is no solution;
	$r(\infty) \leq \bar{r} \leq \lambda_1 b_1$,	$M_1 = \infty$;
	$\lambda_2 b_2 \leq \bar{r} < r(\infty)$, $\lambda_2 > 0$	$M_1 = \lceil x_1 \rceil$;
	$0 \leq \bar{r} < r(\infty)$, $\lambda_2 = 0$	$M_1 = \max(N, \lceil x_1 \rceil)$;
	$\bar{r} < \lambda_2 b_2$	$M_1 = 0$;

where

$$x_1 = -\log(1+y)/\log\rho_1 \quad \text{if } \rho_1 \neq 1,$$
$$= (\bar{r}-\lambda_2 b_2)/((\lambda_1 b_1-\bar{r})(1-\rho_2)) \quad \text{if } \rho_1 = 1,$$
$$y = ((\bar{r}-\lambda_2 b_2)(1-\rho_1))/((\lambda_1 b_1-\bar{r})(1-\rho_2))$$

and $\lceil x \rceil$ denotes the smallest integer greater than or equal to x.

The proof of this theorem is easy and can be found in [3]. Observe that if $\lambda_2=0$, the level M takes on a special meaning: it is the maximum number of customers the server allows in the system. If x_1 as defined in the theorem is less than (N-1), then the solution is not unique, since $P_N(M)=0$ and $r(M)\geq\bar{r}$ for all $M=\lceil x_1 \rceil, \lceil x_1 \rceil+1,\ldots,N$. However, since the critical level has been set to N and not a smaller value, M_1 should be N since $r(N)>r(N-1)>\ldots$ Therefore, if $\lambda_2=0$, then $M_1=\max(N,\lceil x_1 \rceil)$.

If $\lambda_2\neq 0$, then M_1 does not depend on N, therefore, if the objective is to keep the expected fee at the level $\bar{r}$ at least, it is not necessary to determine a critical level N: M_1 minimizes the tail probabilities $P_i(M)$ for all $i>0$ and minimizes the expected queue length. To analyze M_1 as a function of the different parameters, one analyzes either the explicit expression for x_1 or $r(M)$. It appears that M_1 is, trivially, a (non strictly) increasing function of $\bar{r}$ and a decreasing function of all the other parameters: λ_1, λ_2, μ, b_1 and b_2.

One can show also that if $\bar{r}$ is a linear combination of $\lambda_1 b_1$ and $\lambda_2 b_2$,

$$\bar{r} = w\lambda_1 b_1 + (1-w)\lambda_2 b_2$$

for some $w\in(0,1)$, then M_1 is finite if and only if μ is greater than the same linear combination of λ_1 and λ_2:

$$\mu > w\lambda_1 + (1-w)\lambda_2.$$

3. Constraint on the Tail Probability

Suppose now, contrary to the preceding section, that it is decided that the probability of having more than N customers in the system has to be less than some given value ε. The problem is then to find M_2 such that

$$P_N(M_2) \leq \varepsilon$$

and

$$r(M_2) \geq r(M) \quad \text{for all } M \text{ such that } P_N(M) \leq \varepsilon. \tag{4}$$

M_2 is determined very easily. Two cases must be distinguished: if $P_N(N) \leq \varepsilon$, it results from Lemma 1 that M_2 is greater than or equal to N and is determined using equation (1a); if $P_N(N) > \varepsilon$, $M_2 < N$ and equation (1b) has to be used.

Note that

$$P_N(\infty) = \lim_{M \to \infty} P_N(M) = \min(1, \rho_1^{N+1})$$

and that $P_N(0) = \rho_2^{N+1}$ should be smaller than or equal to ε in order to have a solution.

Theorem 3

If $\varepsilon < \rho_2^{N+1}$, there is no solution;

$\rho_2^{N+1} \leq \varepsilon < \min(1, \rho_1^{N+1})$, $M_2 = \lfloor x_2 \rfloor$;

$\min(1, \rho_1^{N+1}) \leq \varepsilon \leq 1$, $M_2 = \infty$;

where $x_2 < N$ is the unique nonnegative root of the equation

$$(\rho_1/\rho_2)^x = \varepsilon(1-\rho_2-\rho_1^x(\rho_1-\rho_2))/(\rho_2^{N+1}(1-\rho_1)) \tag{5}$$

if

$$\rho_2^{N+1} \leq \varepsilon < P_N(N); \tag{6}$$

and

$$x_2 = \log\{(1-\rho_2)(\rho_1^{N+1}-\varepsilon)/((\rho_1-\rho_2)(1-\varepsilon))\}/\log\rho_1 \geq N$$

$$\text{if} \qquad P_N(N) \leq \varepsilon < \min(1, \rho_1^{N+1}) \qquad (7)$$

and $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x. Proof of this theorem can be found in [3].

In the strict sense, equation (5) may have two non negative roots if $\varepsilon = \rho_2^{N+1}$: x=0 and some positive root, say x*. It can be seen easily that in this case, x*<1 and therefore there is no ambiguity on the value of M_2: both $\lfloor 0 \rfloor$ and $\lfloor x^* \rfloor$ are equal to 0.

Observe that x_2 does not depend on b_1 or b_2. Therefore, if the objective is to keep the probability $P_N(M)$ at the level ε at most ($\varepsilon \geq \rho_2^{N+1}$), there is no need to determine precise values for b_1 and b_2; it suffices to check whether $\lambda_1 b_1 > \lambda_2 b_2$ (it is more advantageous to let many customers in the system) or not, in the first case, theorem 3 applies, in the second, theorem 1 and $M_2=0$. There is no explicit form for M_2 if $M_2<N$. However, it is very easy to determine numerically a suitable approximation for x_2 using standard methods; as M_2 is integer valued, one does not need a very high precision on this approximation and it is found rapidly.

By analyzing $P_N(M)$ as a function of the different parameters, one shows that, not surprisingly, M_2 is a decreasing function of λ_1 and λ_2 and an increasing function of μ, N and ε.

4. Control with Hysteresis

We shall examine now the second type of control, where hysteresis is introduced in the arrival process.

Let us assume $\lambda_1 b_1 > \lambda_2 b_2$. If $\lambda_1 b_1 \leq \lambda_2 b_2$, it can be shown that the optimal rule is to impose the fee b_2 always.

Let us assume furthermore that $\gamma \geq 0$ where γ is the cost incurred by the server each time there is a change of fee and

$$M-1 \geq m \geq 0 \tag{8}$$

This last inequality means that the fee is certainly b_1 when the system is empty and the only purpose of this assumption is to get a more uniform presentation of the results.

The determination of the state probabilities and the expected reward is not as straightforward as for the preceding type of control. M. Scott [9] derives the probabilities when the maximum queue size is finite.

In [3], we determine these quantities - when the maximum queue size is infinite - by using a different technique that can be roughly described as follows:

Let $r(m,M)=E[\text{fee collected per unit of time in steady state} | (m,M)]$

and $P_i(m,M)=P[\text{more than } i \text{ customers in the system in steady state} | (m,M)]$.

Obviously, $r(m,M)$ is equal to the expected total fee collected during a busy cycle divided by the expected length of a busy cycle and similarly for $P_i(m,M)$.

To study a busy cycle, we decompose it into homogeneous intervals of time: first, there is an idle period. When a first customer joins the queue, there is a random walk with absorbing

boundaries at 0 and M, during this random walk, the arrival rate is constant and equal to λ_1, the fee is b_1. If absorption is in 0, the busy cycle terminates; if absorption is in M, the server suffers a loss γ and a new random walk is initiated, with absorbing boundary in m, arrival rate λ_2 and fee b_2. Upon absorption in m, there is a cost γ, a new random walk of the first type is initiated and the process repeats itself.

The detailed computation for r(m,M) is given in the appendix; $P_i(m,M)$ is determined in much the same way, as can be found in [3].

Eventually, one gets for $P_i(m,M)$:

a. $0 \le i \le m$

$$P_i(m,M) = \frac{(1-\rho_2)(1-\rho_1^{M-m})\rho_1^{i+1} + (M-m)\rho_1^M(\rho_2-\rho_1)(1-\rho_1)}{(1-\rho_2)(1-\rho_1^{M-m}) + (M-m)\rho_1^M(\rho_2-\rho_1)(1-\rho_1)}$$

b. $m-1 \le i \le M$

$$P_i(m,M) = \frac{\rho_1^{i+1}}{1-\rho_2} \quad x$$

$$\frac{(1-\rho_2)^2(1-\rho_1^{M-i}) + (M-i)\,\rho_1^{M-i-1}(\rho_2-\rho_1)(1-\rho_1)(1-\rho_2) + \rho_1^{M-i-1}(\rho_2-\rho_2^{i-m+1})(1-\rho_1)^2}{(1-\rho_2)(1-\rho_1^{M-m}) + (M-m)\rho_1^M(\rho_2-\rho_1)(1-\rho_1)}$$

c. $M-1 \le i$

$$P_i(m,M) = \frac{\rho_1^M \rho_2^{i-M+1}}{1-\rho_2} \quad \frac{(1-\rho_1)^2(1-\rho_2^{M-m})}{(1-\rho_2)(1-\rho_1^{M-m}) + (M-m)\rho_1^M(\rho_2-\rho_1)(1-\rho_1)}$$

and

$$r(m,M) = \lambda_1 b_1 - \lambda_1 \quad \frac{\rho_1^{M-1}(1-\rho_1)^2(2\gamma(1-\rho_2) + (M-m)(b_1\rho_1 - b_2\rho_2))}{(1-\rho_2)(1-\rho_1^{M-m}) + (M-m)\rho_1^M(\rho_2-\rho_1)(1-\rho_1)}$$

5. Analysis of the Expected Fee and the Tail Probabilities

Lemma 3: For all i, $P_i(m,M)$ is an increasing function of m and M.

This property can easily be justified in the same way as Lemma 1. Note that there is no general relationship between $P_i(m,M)$ and $P_i(m',M')$ if $m<m'$ and $M>M'$.

The behavior of $r(m,M)$ as a function of m and M is more complex. Details can be found in [3] but we shall summarize here the most important properties.

Lemma 4: For any given m, $r(m,M)$ is an increasing function of M.

Lemma 5: For any given M, $r(m,M)$ is a unimodal function of m and we denote by m_M the corresponding maximizing value.

In fact, we determine quantities $\gamma_1(M)$ and $\gamma_2(M)$ such that if $\gamma<\gamma_1(M)$ (the cost γ is small) then $m_M=M-1$: for that value of M the reward is maximum when there is no hysteresis; if $\gamma>\gamma_2(M)$ then $m_M=0$: m should be as far apart from M as possible; if γ lies between $\gamma_1(M)$ and $\gamma_2(M)$, m_M is determined by finding the unique root of some equation.

Remark: $m_{M+1}\geq m_M$ for all $M\geq 1$.

These properties and others give us detailed although not complete information on $r(m,M)$.

6. Constraint on the Expected Fee

The problem is to find (m_1,M_1) such that

$$r(m_1,M_1)\geq\bar{r}$$

and $P_N(m_1,M_1)\leq P_N(m,M)$ for all (m,M) such that $r(m,M)\geq\bar{r}$.

Let Ω be the set $\{(m,M)\,|\,M=1,2,\ldots;\ m=0,1,\ldots,M-1\}$ of all pairs satisfying (8) and let

$$\Omega_1 = \{(m,M)\,|\,(m,M)\in\Omega,\ r(m,M)\geq\bar{r}\}.$$

Obviously, $(m_1,M_1)\in\Omega_1$.

If Ω_1 is empty, the problem has no solution; if $\Omega_1=\Omega$, it results from Lemma 3 that $(m_1,M_1)=(0,1)$, the fee should be b_2 whenever there is a customer in the system; otherwise, we proceed in two steps to determine (m_1,M_1).

First, by using repeatedly the properties of $r(m,M)$, we prove that Ω_1 has the form indicated on figure 1: the dashed region corresponds to $\Omega-\Omega_1$; the boundary of Ω_1 is made up of two parts: for increasing m, it is first non-increasing and second it is non-decreasing until it coincides with the line $m=M-1$. We have marked by an x the "corners" in the first part of the boundary. We denote those pairs by (m_i^x,M_i^x). We shall not give here the full definition of the (m_i^x,M_i^x)'s, as it is very cumbersome. Let us mention that they are ordered in such a way that $M_i^x<M_{i+1}^x$ and $m_i^x>m_{i+1}^x$. Their most important properties are:

$$M_1^x=\min\{M\,|\,r(m_M,M)\geq\bar{r}\}\ ;$$

for a given $m=m_i^x$,

$$r(m_i^x,M)\geq\bar{r}\ \text{ iff }\ M\geq M_i^x\ ;$$

for a given $M=M_i^x$,

$$r(m,M_i^x)<\bar{r} \text{ for all } m<m_i^x.$$

We then use Lemma 3 to prove the following theorem:

Theorem 4

$$(m_1,M_1)\in\{(m_i^x,M_i^x),\ i=1,\ldots,I^x\}.$$

There is no explicit expression for (m_1,M_1), one should compare all (m_i^x,M_i^x)'s to determine it. This can easily be done numerically. Note that I^x is at most equal to $\lceil x_1 \rceil$ (defined in Theorem 2) and is, in general, smaller; we have thus an upper bound on the number of pairs that will have to be determined and compared.

To analyze the solution as a function of the parameters, one has to resolve to computations. As an example, we have examined how (m_1,M_1) depends on N (see figure 2). For the single critical number policies, we have seen that M_1 is independent of N. For control with hysteresis, it appears that (m_1,M_1) varies in general only slightly with N, if at all; however, if $\bar{r}$ is very close to the $\lim_{M\to\infty} r(m,M)$ and λ_1 is much larger than λ_2, the variations are greater, as can be seen in figure 2. Those cases do not seem typical.

One can observe that if $\gamma=0$, m_1 is not necessarily equal to (M_1-1) (Table I). In other words, even if the server suffers no cost each time the fee is changed, it is optimal in some cases to introduce hysteresis in the arrival process.

λ_2	.05	.1	.2	.3
m_1	2	2	3	3
M_1	4	4	4	5

$\mu=1.$, $\lambda_1=.9$, $\bar{r}=.8$, $b_1=1.$,

$\gamma=0$, $b_2=.4/\lambda_2$, $N=0,1,\ldots,10$

Table I

7. Constraint on the Tail Probability

(m_2,M_2) is defined as follows

$$P_N(m_2,M_2)\leq\varepsilon$$

and $r(m_2,M_2)\geq r(m,M)$ for all (m,M) such that $P_N(m,M)\leq\varepsilon$.

Let $\Omega_2=\{(m,M)\,|\,(m,M)\in\Omega \text{ and } P_N(m,M)\leq\varepsilon\}$. If Ω_2 is empty, the problem has no solution. If $\Omega_2=\Omega$, it results from Lemma 4 that $M_2=\infty$ and therefore m_2 can take any value: the fee is b_1 always. Otherwise, we use Lemma 3 to prove that Ω_2 has the form indicated on figure 3: the dashed region corresponds to $\Omega-\Omega_2$; the boundary of Ω_2 forms a non-increasing line for increasing m. We have marked by a + some of the points in Ω_2, denoted by (m_i^+,M_i^+). Again, we do not define formally those pairs in this paper. Let us mention that they are ordered in such a way that $M_i^+<M_{i+1}^+$ and $m_i^+>m_{i+1}^+$. (m_1^+,M_1^+) is defined as follows:

$$M_1^+ = \max\{M\,|\,P_N(m_M,M)\leq\varepsilon\},$$

$$m_1^+ = m_{M_1^+}\ ;$$

the other pairs represent the "corners" in the boundary of Ω_2, to the left of (m_1^+,M_1^+); moreover, for a given $m=m_i^+$,

$$P_N(m_i^+,M)\leq\varepsilon \quad \text{iff} \quad M\leq M_i^+$$

and for a given $M=M_i^+$,

$$P_N(m,M_i^+)\leq\varepsilon \quad \text{iff} \quad m\leq m_i^+.$$

We then use the properties of r(m,M) to prove the following theorem:

Theorem 5:

$$(m_2,M_2)\in\{(m_i^+,M_i^+),\ i=1,\ldots,I^+\}\ .$$

To complete theorem 5, we may add that

$$\text{if } (m_1^+,M_1^+-1)\in\Omega_2,\ \text{then } (m_2,M_2)\neq(m_1^+,M_1^+)\ .$$

In other words, (m_2,M_2) lies on the boundary of Ω_2 and (m_1^+,M_1^+) cannot be optimal if it lies inside Ω_2. x_2 (defined in Theorem 3) provides an upper bound on the number of pairs that have to be determined and compared, for $I^+\leq(\lfloor x_2\rfloor+1)$.

We have analyzed numerically (m_2,M_2) as a function of γ and it appears (figures 4 to 6) that M_2 is a (non strictly) increasing function of γ while m_2 is decreasing. This property is not surprising: if the cost γ increases, the server should wait longer before switching to the fee b_2 (M_2 increases) and wait longer before switching back to the fee b_1 (the difference (M_2-m_2) increases). This shows that, to the contrary of the single critical number policies (see end of Section 3), the optimal solution depends on the values assigned to cost and fees.

Observe that if $\gamma=0$, m_2 is not necessarily equal to (M_2-1) (figures 4 and 6).

8. Comparisons Between the Two Types of Policies

To conclude, we mention numerical comparisons between the two types of policies when the cost γ is positive: we can determine as in Sections 1 to 3 optimal single critical number policies.

Figures 7 to 10 present 4 examples corresponding to 2 values for N (N=5 and N=10) and 2 values for γ (γ=.5 - which is small compared to b_1 and b_2 - and γ=5 - which is large compared to b_1 and b_2). On each figure, curves 1 and 2 correspond to γ=.5, curves 3 and 4 to γ=5; curves 1 and 3 correspond to control with hysteresis, curves 2 and 4 to control with a single critical number.

Obviously, controls with a single critical number yields less good results, since they are special cases of controls with hysteresis. As can be seen in the figures, the difference can be substantial. Also, it appears that the optimal control with hysteresis is efficient, except under severe circumstances, such as small value for N, large cost γ and quite large $\bar{r}$ (or small ε).

Appendix

Let us denote by $r_{m,M}$ the expected total reward during a busy cycle given m and M. Consider a random walk on $\{0,1,\ldots,L\}$ with absorbing boundaries at 0 and L. Transitions are from n to n+1 (step to the right) and n to n-1 (step to the left) if $n\neq 0$ and L. The interval of time between two steps to the right (to the left) is negative exponential with parameter $\lambda(\mu)$. Because of the memoryless property of the negative exponential distribution, the interval of time between any step and a step to the right (left) is negative exponential with parameter $\lambda(\mu)$. Moreover, $P[n\to n+1]=\lambda/(\lambda+\mu)$ and $P[n\to n-1]=\mu/(\lambda+\mu)$ ($n\neq 0$ and L). Let

$a_n(L,\lambda,\mu)=P[\text{absorption in 0}\,|\,\text{initial state is n}]$

$g_n(L,\lambda,\mu)=E[\text{number of steps to the right before absorption in 0 or L}\,|\,\text{initial state is n}]$.

It is well known, see for instance [2] p. 314 that

$$a_n(L,\lambda,\mu)=(1-\rho^L)^{-1}(1-\rho^{L-n}).$$

where $\rho=\lambda/\mu$.

We show in [3] that

$$g_n(L,\lambda,\mu)=(1-\rho)^{-1}\rho[n-L(1-a_n(L,\lambda,\mu))].$$

If L tends to ∞, we get, if $\rho<1$,

$$a_n(\infty,\lambda,\mu)=1$$

$$g_n(\infty,\lambda,\mu)=n\rho(1-\rho)^{-1}.$$

To determine $r_{m,M}$, we decompose a busy cycle as follows. First, there is an idle period, at the end of which a customer enters the queue and pays b_1. Then a random walk begins, with absorbing boundary at 0 and M, parameter λ_1, to each step to the

right is associated a reward b_1. If absorption is at 0, the busy cycle terminates. Otherwise, a cost γ is incurred and a new random walk begins, with absorbing boundary at m, parameter λ_2, reward b_2 for each step to the right. Upon absorption at m, a cost γ is incurred and a new random walk of the first type is initiated, etc... Thus, if $r_m(r_M)$ is the total expected reward during the remainder of the busy cycle given that the queue is in state m(M) and if $g_{0,n}(L,\lambda,\mu)$ (resp. $g_{L,n}(L,\lambda,\mu)$) is the expected number of steps to the right before absorption given that the initial state is n and absorption is in 0 (resp. L) we get:

$$r_{m,M}=b_1+b_1a_1(M,\lambda_1,\mu)g_{0,1}(M,\lambda_1,\mu)+(1-a_1(M,\lambda_1,\mu))(b_1g_{M,1}(M,\lambda_1,\mu)+r_M-\gamma),$$

$$r_M=b_2g_{M-m}(\infty,\lambda_2,\mu)+r_m-\gamma,$$

$$r_m=b_1a_m(M,\lambda_1,\mu)g_{0,m}(M,\lambda_1,\mu)+(1-a_m(M,\lambda_1,\mu))(b_1g_{M,m}(M,\lambda_1,\mu)+r_M-\gamma).$$

As $g_n=a_ng_{0,n}+(1-a_n)g_{L,n}$, one gets eventually

$$r_{m,M}=b_1\left[\frac{1}{1-\rho_1}-\frac{(M-m)\rho_1^M(\rho_1-\rho_2)}{(1-\rho_2)(1-\rho_1^{M-m})}\right]-\frac{\rho_1^{M-1}(1-\rho_1)[2\gamma(1-\rho_2)+(M-m)(b_1\rho_1-b_2\rho_2)]}{(1-\rho_2)(1-\rho_1^{M-m})}. \qquad (A.1)$$

Let $\omega_{m,M}$ be the expected length of a busy cycle. We determine $\omega_{m,M}$ in [3] by the same technique we used to determine $r_{m,M}$. We propose here another argument: $\omega_{m,M}$ is the idle period (mean $1/\lambda_1$) plus the amount of work brought by the customers during the busy cycle. To get this amount of work, one just replaces in (A.1) γ by 0, b_1 and b_2 by $1/\mu$.

This leads to

$$\omega_{m,M}=\frac{1}{\lambda_1}\left[\frac{1}{1-\rho_1}-\frac{(M-m)\rho_1^M(\rho_1-\rho_2)}{(1-\rho_2)(1-\rho_1^{M-m})}\right] \qquad (A.2)$$

From (A.1) and (A.2), it is easy to check that

$$r(m,M)=\frac{r_{m,M}}{\omega_{m,M}}$$

$$=\lambda_1 b_1-\lambda_1\,\frac{\rho_1^{M-1}(1-\rho_1)^2(2\gamma(1-\rho_2)+(M-m)(b_1\rho_1-b_2\rho_2))}{(1-\rho_2)(1-\rho_1^{M-m})+(M-m)\rho_1^M(\rho_2-\rho_1)(1-\rho_1)}$$

BIBLIOGRAPHY

[1] N. Edelson and D. Hildebrand, "Congestion Tolls for Poisson Queueing Processes", Econometrica 43, 81-92 (1975).

[2] W. Feller, "An Introduction to Probability Theory and its Applications", Vol. 1, J. Wiley & Sons (1957).

[3] G. Latouche, "Modèles de Contrôle Optimal d'une File d'Attente avec Droit d'Entrée Variable", Ph.D. Dissertation, Faculté des Sciences, Université Libre de Bruxelles, Brussels (1976).

[4] G. Latouche, "Optimal pricing policy for a M/M/1 queue a) without and b) with hysteresis in the arrival process", Proceedings of the Second European Congress on Operations Research, Stockholm 1976, North-Holland, 251-258.

[5] D. Low, "Optimal Dynamic Operating Policies for an M/M/s Queue with Variable Arrival Rate", IBM Los Angeles Scientific Center, G320-2654 (1971).

[6] D. Low, "Optimal Dynamic Pricing Policies for an M/M/s Queue", Operations Research 22, 545-561 (1974).

[7] P. Naor, "The Regulation of Queue Size by Levying Tolls", Econometrica 37, 15-24 (1969).

[8] N. Nielsen, "Flexible Pricing: An Approach to the Allocation of Computer Resources", AFIPS Conference Proceedings 33, 521-531 (1968).

[9] M. Scott, "A Queueing Process with Some Discrimination", Management Science 16, 227-233 (1969).

[10] J. Teghem Jr., "Comportement Optimal des Clients dans Quelques Modèles d'Attente", Ph.D. Dissertation, Faculté des Sciences, Université Libre de Bruxelles, Brussels (1976).

[11] M. Yadin & P. Naor, "On Queueing Systems with Variable Service Capacities", Naval Research Logistic Quarterly 14, 43-53 (1967).

[12] U. Yechiali, "On Optimal Balking Rules and Toll Charges in the GI/M/1 Queueing Process", Operations Research 19, 349-370 (1971).

[13] U. Yechiali, "Customer's Optimal Joining Rules for the GI/M/s Queue", Management Science 18, 434-443 (1972).

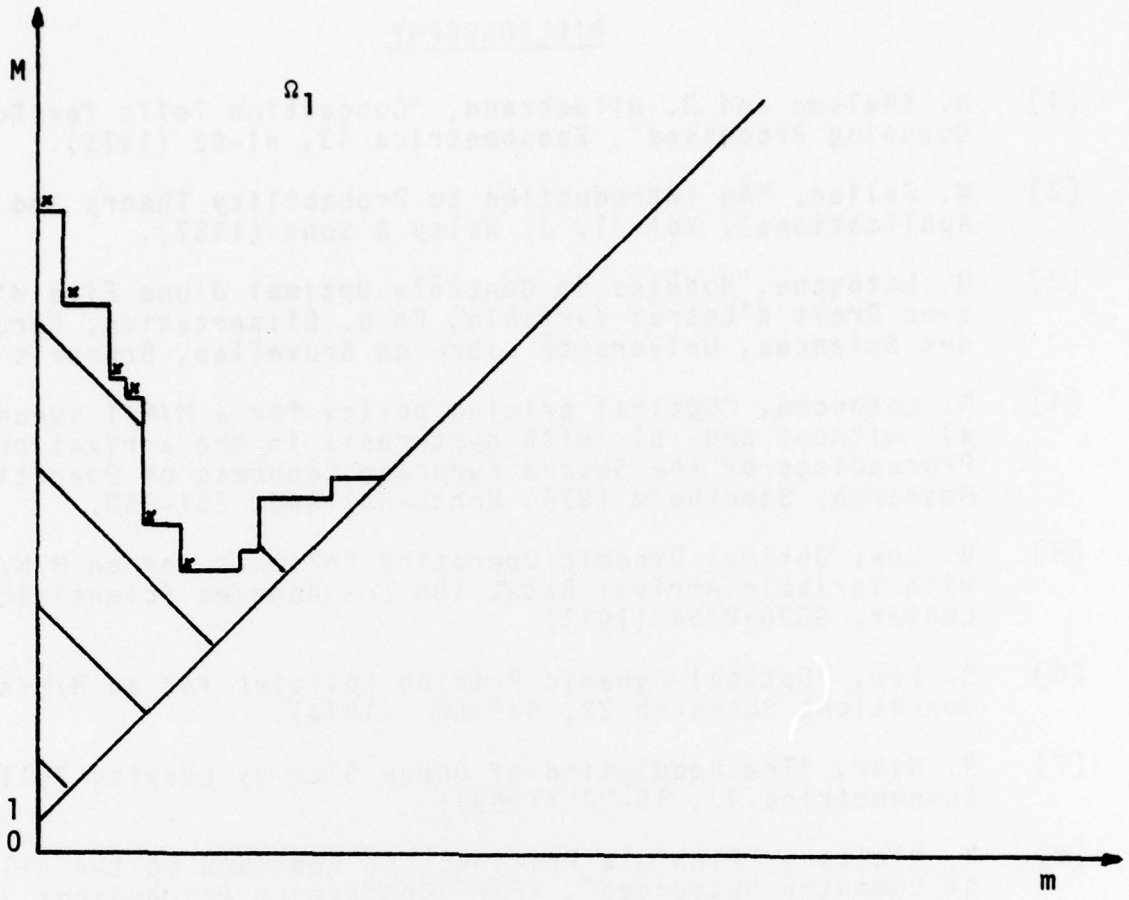
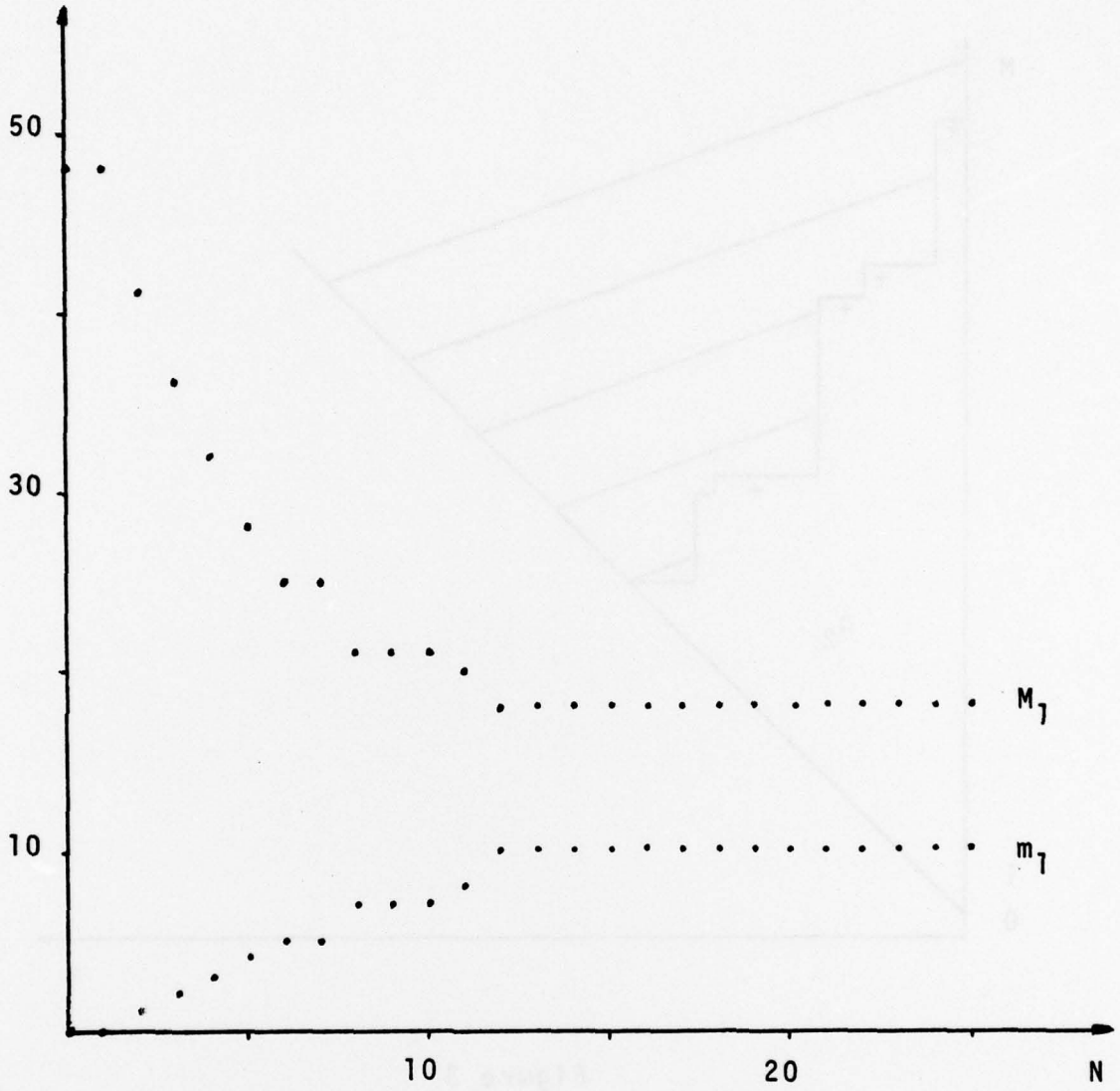


Figure 1



$\mu=1.$, $\lambda_1=1.2$, $\lambda_2=.1$, $b=1$, $B=4$, $\gamma=.0625$, $\bar{r}=.8$, $r(\infty)=.809$

Figure 2

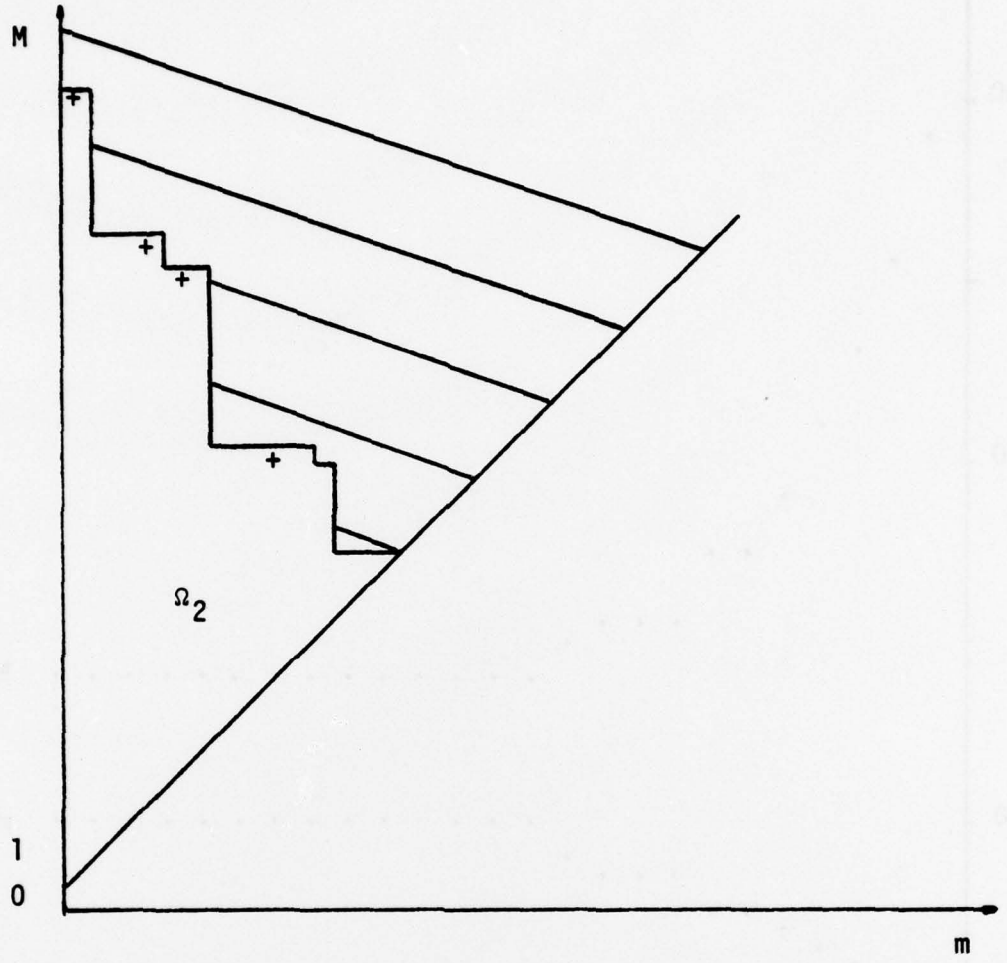
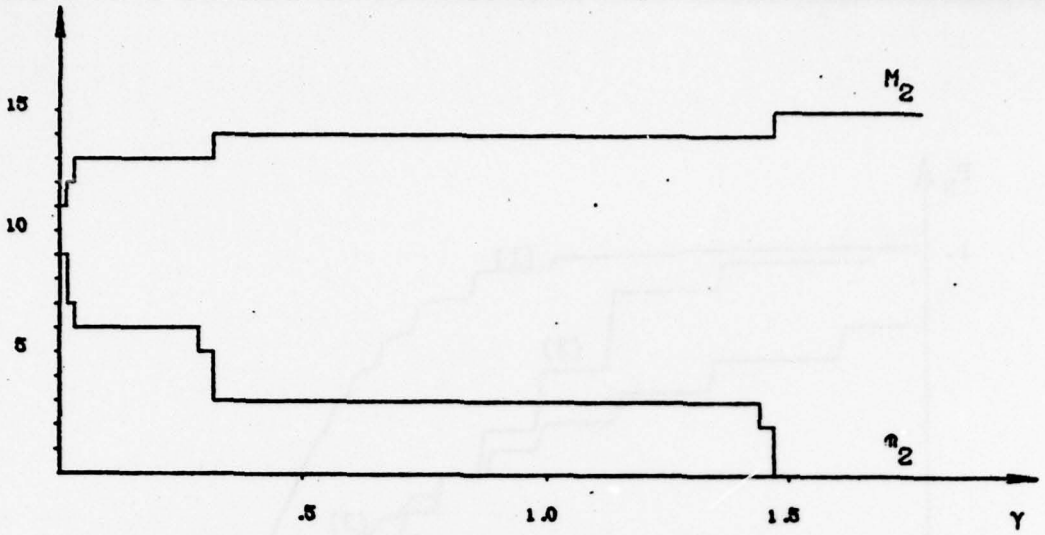
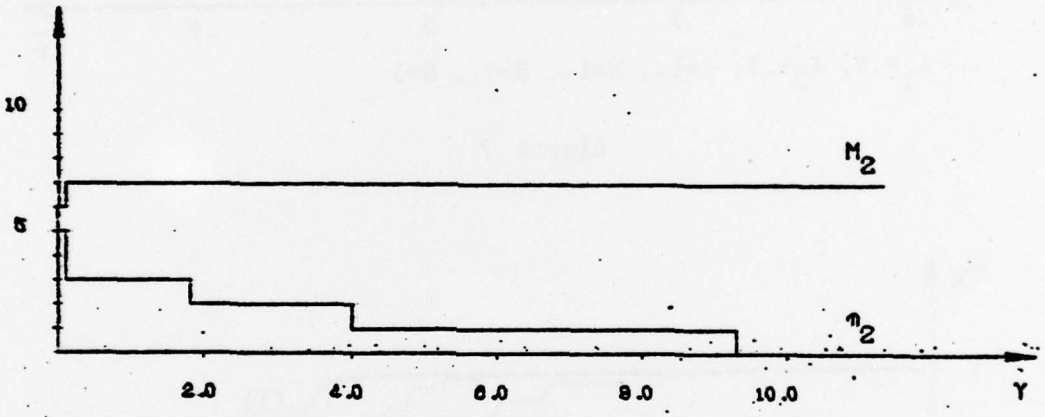


Figure 3



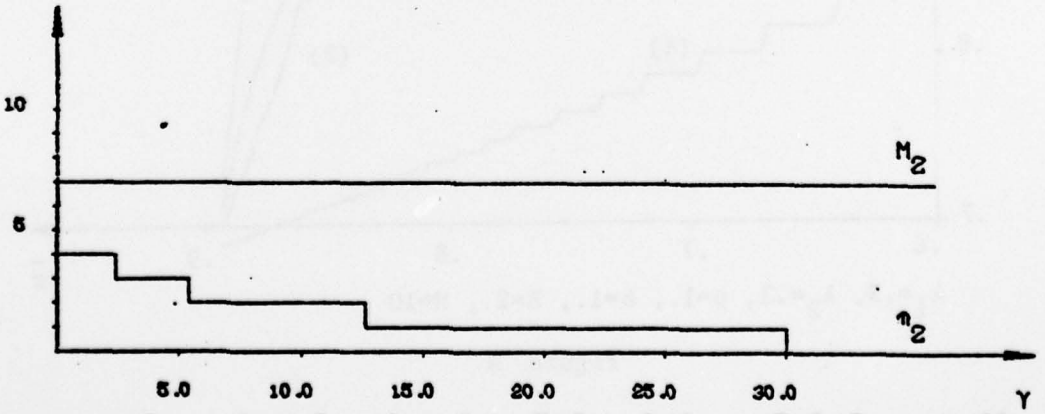
μ=1. , λ_1=1.3 , λ_2=.4 , N=10 , b=1. , B=2. , ε=.25

figure 4



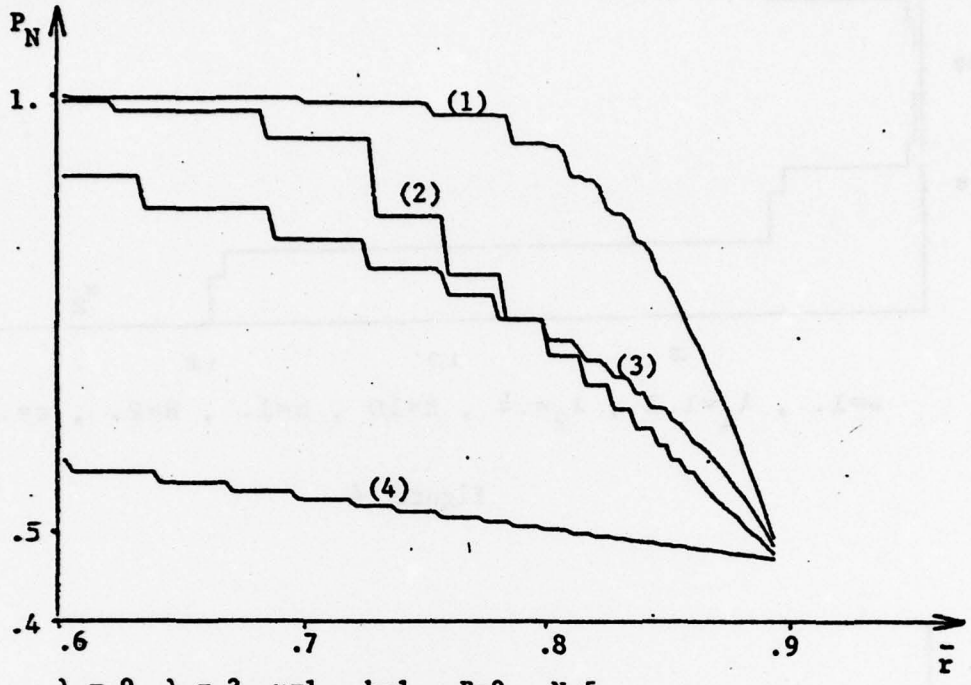
μ=1. , λ_1=.5 , λ_2=.2 , N=5 , b=1. , B=2. , ε=.01

figure 5



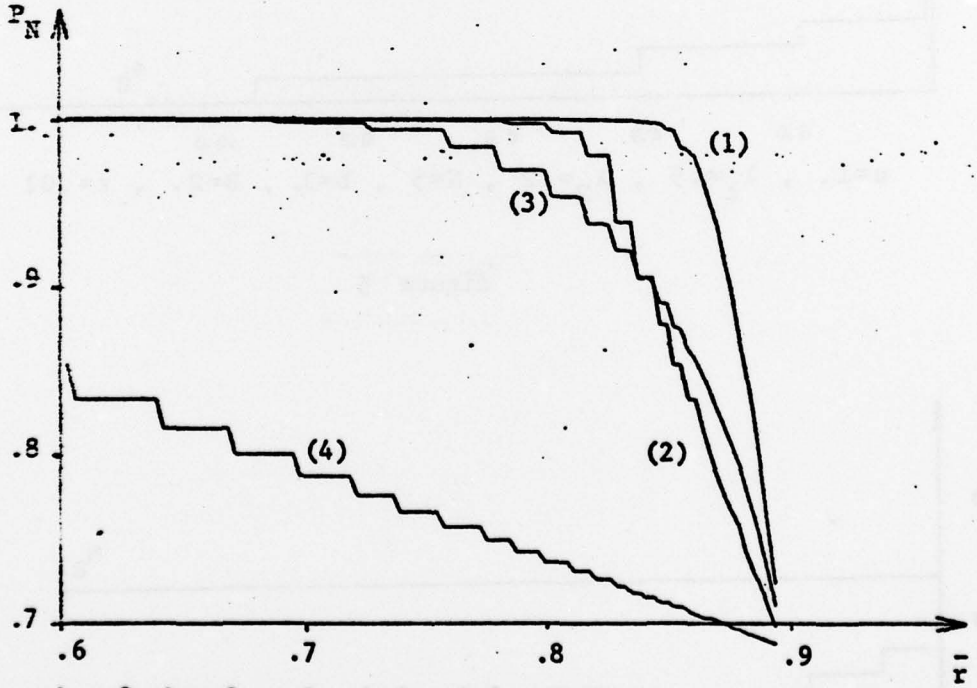
μ=1. , λ_1=.5 , λ_2=.1 , N=5 , b=1. , B=1.5 , ε=.01

figure 6



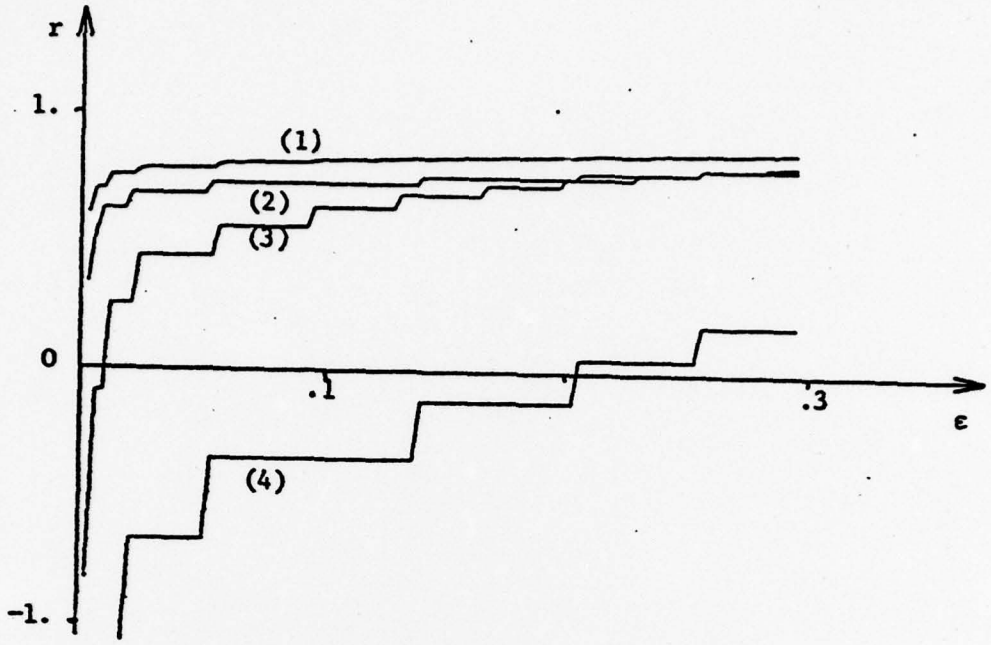
λ_1=.9, λ_2=.3, μ=1., b=1., B=2., N=5

figure 7



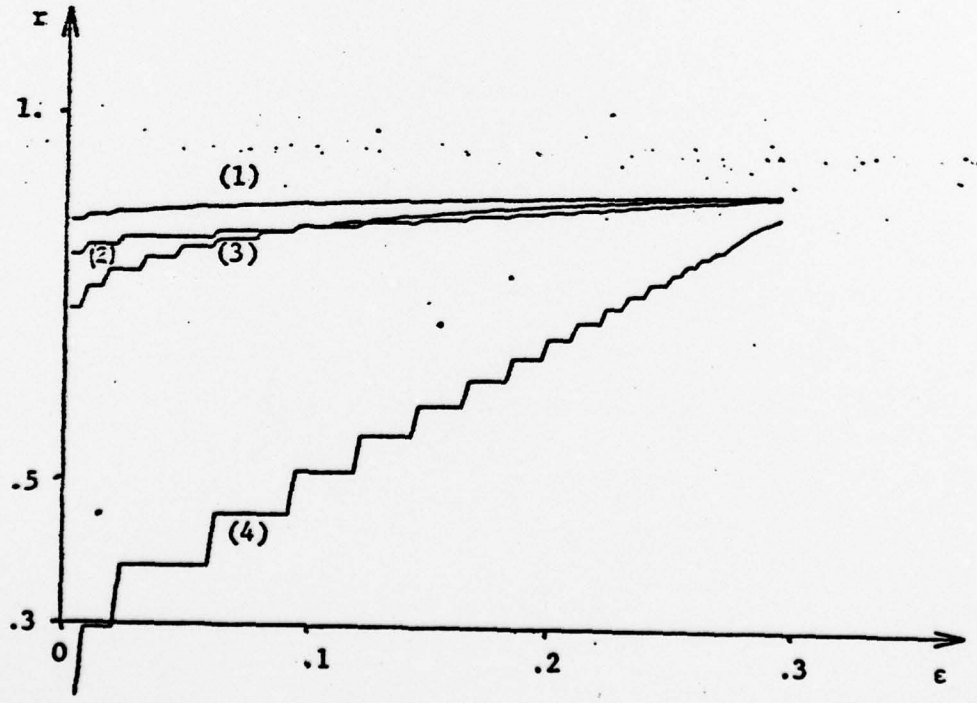
λ_1=.9, λ_2=.3, μ=1., b=1., B=2., N=10

figure 8



λ_1=.9, λ_2=.3, μ=1., b=1., B=2., N=5

figure 9



λ_1=.9, λ_2=.3, μ=1., b=1., B=2., N=10

figure 10

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR-TR- 78-0985 / 2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) ON THE TRADE-OFF BETWEEN QUEUE CONGESTION AND SERVER'S REWARD IN A M/M/1 QUEUE	5. TYPE OF REPORT & PERIOD COVERED Interim
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Guy Latouche	8. CONTRACT OR GRANT NUMBER(s) AFOSR-77-3236
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Delaware Dept. of Statistics & Computer Science Newark, DE 19711	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2304/A5
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research/NM Bolling AFB, Washington, DC 20332	12. REPORT DATE April 1978
	13. NUMBER OF PAGES 31
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report) UNCLASSFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We consider an M/M/1 queue which is controlled by dynamically setting the customer's entrance fee to either b_1 or $b_2 > b_1$ and thereby setting the customer arrival rate to either λ_1 or $\lambda_2 < \lambda_1$, respectively. With E[F]

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. Abstract

the expected fee collected per unit of time and P_N the steady state probability that the system contains more than N customers, we consider two criteria: (i) for some number $\bar{r}$, minimize P_N subject to $E[F] \geq \bar{r}$, and (ii) for some number ε, maximize $E[F]$ subject to $P_N \leq \varepsilon$. Under each criterion we consider two cases: (a) the admissible policies are single critical number switching policies and (b) a cost $\gamma \geq 0$ is incurred whenever the server switches between b_1 and b_2, the admissible policies allow for hysteresis to appear in the arrival process. Optimal policies are computed for each criterion and each case.